Numer. Math. (2016) 133:819–843
DOI 10.1007/s00211-015-0759-9

Numerische Mathematik

Backward error analysis of the shift-and-invert Arnoldi algorithm

Christian Schröder [1] · **Leo Taslaman**[2]

Received: 29 October 2014 / Revised: 29 May 2015 / Published online: 30 July 2015

Abstract We perform a backward error analysis of the inexact shift-and-invert Arnoldi algorithm. We consider inexactness in the solution of the arising linear systems, as well as in the orthonormalization steps, and take the non-orthonormality of the computed Krylov basis into account. We show that the computed basis and Hessenberg matrix satisfy an exact shift-and-invert Krylov relation for a perturbed matrix, and we give bounds for the perturbation. We show that the shift-and-invert Arnoldi algorithm is backward stable if the condition number of the small Hessenberg matrix is not too large. This condition is then relaxed using implicit restarts. Moreover, we give notes on the Hermitian case, considering Hermitian backward errors, and finally, we use our analysis to derive a sensible breakdown condition.

Mathematics Subject Classification 65F25 · 65F50 · 65G50

1 Introduction

Consider an implementation of the Arnoldi algorithm [4,26]. Not much meaning can be given to the computed quantities if they deviate too much from the recurrence that underpins the algorithm in exact arithmetic:

C. Schröder was supported by German research council, DFG under project "Scalable Numerical Methods for Adiabatic Quantum Preparation". L. Taslaman was supported by Engineering and Physical Sciences Research Council Grant EP/I005293.

✉ Leo Taslaman
leotaslaman@gmail.com

Christian Schröder
schroed@math.tu-berlin.de

[1] Institut für Mathematik, MA 4–5, Technische Universität Berlin, Berlin, Germany

[2] School of Mathematics, The University of Manchester, Manchester M13 9PL, UK

$$AV_k = V_{k+1}\underline{H}_k, \quad \underline{H}_k = H(1:k+1, 1:k).$$

Luckily, good implementations, where in particular the orthogonalization is done with care, can be shown to be backward stable [3,8,10,21] in the sense that the computed quantities V_{k+1} and $\underline{H}_k$ satisfy an exact recurrence with a slightly perturbed matrix:

$$(A + \Delta A)V_k = V_{k+1}\underline{H}_k. \tag{1}$$

This means that we can compute a basis of an exact Krylov subspace corresponding to a nearby matrix. Since the basis will in general not be perfectly orthonormal, so $V_{k+1}^H V_{k+1} \neq I$, we use the term "Krylov recurrence" instead of "Arnoldi recurrence" when referring to recurrences like (1), as suggested in [24]. If A is Hermitian, then it can be shown that the computed basis spans a Krylov subspace associated with a perturbed *Hermitian* matrix $A + \Delta A$ [15]. There is a catch in this case, though: the small $(k+1) \times k$ matrix associated with this Krylov subspace is in general not the computed Hessenberg matrix.

In this paper we perform a similar backward error analysis of the shift-and-invert Arnoldi algorithm. For example, we show that an implementation of the Arnoldi algorithm applied to A^{-1}, yields computed matrices V_{k+1} and $\underline{H}_k$ such that

$$(A + \Delta A)^{-1}V_k = V_{k+1}\underline{H}_k,$$

and we give an upper bound for $\|\Delta A\|_2$. Perturbed versions of the shift-and-invert Arnoldi algorithm have been considered in the literature as a part of the theory of *inexact methods*, see [16,19]. However, these results neglect that the orthonormalization is not performed exactly, and furthermore, assume bounds on linear system residuals that may be unattainable (more on this in Sect. 2). We consider more general linear system residuals and take the error from the orthonormalization into account. Our analysis of how the orthonormalization errors propagate into the shift-and-invert Krylov recurrence highlights the importance of *columnwise* backward error bounds for QR factorizations, and is thus of a different flavor than the corresponding analysis for standard Arnoldi, done in, for example [8].

We also use our error analysis to motivate when "breakdown" should be declared, that is, when $h_{j+1,j}$ may be considered to be "numerically zero".

The algorithm we study can be divided into two main subproblems: solving linear systems and orthonormalizing vectors. We state our backward error results in such a way that they are independent of how these subproblems are being solved, but we also discuss relevant and commonly used approaches for solving these two tasks.

1.1 Technical outline

We study floating point implementations of Algorithm 1, where A is assumed to be of size $n \times n$, σ is the shift, b the starting vector, and k is the maximum number of steps we perform. Throughout the paper $\|\cdot\|$ refers to the 2-norm.

Algorithm 1 The Shift-and-invert Arnoldi algorithm

Input: A, σ, b, k
Output: $V_{k+1} = [v_1\ v_2\ \cdots\ v_{k+1}]$, $\underline{H_k} = [h_{ij}]_{i=1:k+1, j=1:k}$
$v_1 = b/\|b\|$
for $j = 1, 2, \ldots, k$
$w_j = (A - \sigma I)^{-1} v_j$
$[w'_j, h_{1:j}] = \text{orthogonalization}(w_j, V_j)$
$h_{j+1,\,j} = \|w'_j\|$
if $h_{j+1\,j} = 0$ **break**
$v_{j+1} = w'_j / h_{j+1\,j}$
end for

In exact arithmetic, we have

$$\text{orthogonalization}(w_j, V_j) := \left[w_j - V_j \left(V_j^H w_j \right), V_j^H w_j \right],$$

which corresponds to classical Gram–Schmidt if implemented as it stands. In floating point arithmetic, orthogonalization routines with better numerical properties, such as modified Gram–Schmidt (MGS), are usually employed.

In the jth iteration in Algorithm 1, a new vector w_j is computed and decomposed into a linear combination of $v_1, \ldots, v_j$ and a new component that will be the definition of v_{j+1}. In exact arithmetic, this can be described by the Arnoldi recurrence

$$(A - \sigma I)^{-1} v_j = V_k h_{1:j,j} + h_{j+1,j} v_{j+1}.$$

When the corresponding thing is done in practice, however, errors are present in all steps of the computation. First, we need to solve a linear system. If we use a direct solver the matrix $A - \sigma I$ needs to be formed. We consider the rounding error in this step as part of the residual from the linear system. This does not affect the norm of the residual significantly, because the rounding error is very small,

$$\|\text{float}(A - \sigma I) - (A - \sigma I)\| < \max_{1 \le i \le n} |a_{ii} - \sigma| u \le u \|A - \sigma I\|.$$

Here $\text{float}(A - \sigma I)$ refers to the computed shifted matrix and u is the unit roundoff. Let r_j be the said residual from the linear system, so

$$(A - \sigma I) w_j = v_j + r_j \tag{2}$$

is the actual linear system that has been solved. Then we have the following equality for the computed quantities:

$$(A - \sigma I)^{-1} (v_j + r_j) = w_j = V_{j+1} h_{1:j+1,j} + g_j,$$

where g_j is an error coming from the orthonormalization process. Defining

$$f_j = r_j - (A - \sigma I) g_j$$

and $F_k = [f_1\ f_2\ \cdots\ f_k]$ yields a perturbed recurrence

$$(A - \sigma I)^{-1}(V_k + F_k) = V_{j+1}\underline{H}_k.$$

We discuss the residual r_j and the error g_j in Sects. 2 and 3, respectively, and provide bounds for both quantities. In Sect. 4, we use these bounds in order to bound F_k, and subsequently the backward error for the shift-and-invert Arnoldi recurrence. In Sect. 5, we explain how the idea of implicit restarting can be used to gain further insight into the backward error. We also discuss in what sense we have Hermitian backward errors if the method is applied to a Hermitian matrix A. Finally, we talk about breakdown conditions: in floating point arithmetic, the test if $h_{j+1,j} = 0$ in Algorithm 1 is rarely done. Instead one usually checks whether $h_{j+1,j}$ is "small enough". This case is referred to as *breakdown*. A sensible definition of "small enough" is when the quantity is dominated by errors. We discuss this in more detail and derive backward error bounds for this case.

1.2 Notation

The scalar σ refers to a shift while $\sigma_{\min}(X)$ refers to the smallest singular value of X. The dagger notation $X^\dagger$ refers to the Moore-Penrose pseudo-inverse of X. The lower letter u is reserved to denote the unit roundoff if real arithmetic is used, and $\sqrt{5}$ times the unit roundoff if complex arithmetic is used [7]. When the matrix size is understood from the context, we denote zero matrices and identity matrices as 0 and I, respectively. Similarly, the vector e_i denotes the ith column of the identity matrix whose size is understood from the context. For a matrix X, the lower case x_i refers to the ith column of X and X_k to $[x_1\ x_2\ \cdots\ x_k]$, that is, the first k columns of X.

2 Errors from linear systems

In this section we discuss bounds on the residual r_j from (2).

2.1 Backward error bounds

The normwise backward error associated with a computed solution y of a linear system $Ax = b$ is defined as

$$\eta_{A,b}(y) := \min\{\epsilon : (A + \Delta A)y = b + \Delta b, \|\Delta A\| \le \epsilon\|A\|, \|\Delta b\| \le \epsilon\|b\|\},$$

and given by the formula

$$\eta_{A,b}(y) = \|r\|/(\|A\|\|y\| + \|b\|) \tag{3}$$

where $r = Ay - b$ [20]. See also [12, p. 120]. This result is true for any vector norm $\|\cdot\|$ and its subordinate matrix norm. Thus, if we solve the linear systems in Algorithm 1, up to a backward error ϵ_{bw}, then it holds that

$$\|r_j\| \leq (\|A - \sigma I\|\|w_j\| + \|v_j\|)\epsilon_{\text{bw}}, \tag{4}$$

where r_j is defined in (2). If the linear systems are solved by a backward stable direct method, we have $\epsilon_{\text{bw}} \leq \phi(n)u$, where $\phi(n)$ is an algorithm dependent constant. If we are interested in the smallest possible ϵ_{bw} such that (4) holds, then we need to compute $\|r_j\|/(\|A - \sigma I\|\|w_j\| + \|v_j\|)$. However, this may not be feasible for the 2-norm, due to the term $\|A - \sigma I\|$. In these cases we can replace $\|A - \sigma I\|$ by a lower bound (the tighter the better), and thus obtain an upper bound for ϵ_{bw}. We can for instance do a few iterations of the power method applied to $(A - \sigma I)^H(A - \sigma I)$. MATLAB's `normest` function does exactly this. This would lead to a lower bound of $\|A - \sigma I\|$, since convergence is always from below. Another possibility is to use the (lower) bound in [13]. We can also bound the matrix 2-norm in terms of the corresponding infinity-norm or 1-norm. The following proposition shows that such bounds can be satisfactory for many sparse matrices, in particular those which can be permuted to banded form.

Proposition 1 *Let k_{row} and k_{col} denote the maximum number of nonzero entries in a row and column of A, respectively. Then the following two upper and lower bounds hold:*

$$\begin{aligned} \frac{1}{\sqrt{k_{\text{col}}}}\|A\|_2 \;&\leq\; \|A\|_\infty \;\leq\; \sqrt{k_{\text{row}}}\|A\|_2, \\ \frac{1}{\sqrt{k_{\text{row}}}}\|A\|_2 \;&\leq\; \|A\|_1 \;\leq\; \sqrt{k_{\text{col}}}\|A\|_2. \end{aligned}$$

Proof We have $\|A\|_\infty = \|Ax\|_\infty$ for some x with $\|x\|_\infty = 1$ and at most k_{row} nonzeros. We get

$$\|A\|_\infty \leq \|Ax\|_\infty \leq \|Ax\|_2 \leq \|A\|_2\|x\|_2 \leq \sqrt{k_{\text{row}}}\|A\|_2,$$

which is the desired upper bound for $\|A\|_\infty$. Further, we have

$$\|A\|_1 = \|A^T\|_\infty \leq \sqrt{k_{\text{col}}}\|A^T\|_2 = \sqrt{k_{\text{col}}}\|A\|_2,$$

which is the desired upper bound for $\|A\|_1$.

The lower bounds follow from [22, Theorem 4.2]. □

The inequality (4) can also be used as a stopping criterion for iterative linear system solvers [2]. In this case, ϵ_{bw} denotes the desired backward error, which is given prior to execution. If we replace $\|A - \sigma I\|$ with a lower bound, then we get a more stringent stopping criterion.

2.2 Residual reduction bounds

An alternative to (4) is to use the bound

$$\|r_j\| \leq \|v_j\|\epsilon_{\text{tol}}. \tag{5}$$

This bound is commonly used as a stopping condition when the linear systems are solved by iterative methods. Unfortunately, as a stopping condition, (5) "may be very stringent, and possibly unsatisfiable" [12, p. 336]. See also [9, pp. 72–73] for a 2×2 example that illustrates the pitfall of comparing the norm of the residual with the norm of the right hand side. However, since (5) is de facto commonly used in computer codes it is still worth to study it under the assumption that the stopping criterion is met.

2.3 Auxiliary residual bounds

In order to treat both (4) and (5) in a unified way, we consider the following auxiliary bound

$$\|r_j\| \leq \|v_j\|\epsilon_1 + \|A - \sigma I\|\|w_j\|\epsilon_2. \tag{6}$$

Clearly, the substitutions $(\epsilon_1, \epsilon_2) \leftarrow (\epsilon_{\text{bw}}, \epsilon_{\text{bw}})$ and $(\epsilon_1, \epsilon_2) \leftarrow (\epsilon_{\text{tol}}, 0)$ give back (4) and (5), respectively. We can simplify the bound in (6) in cases when $A - \sigma I$ is not too ill-conditioned with respect to ϵ_2. To see this we need the following lemma.

Lemma 2 *If* $\kappa(A - \sigma I)\epsilon_2 < 1$ *and* (6) *hold, then*

$$\|r_j\| \leq \frac{\epsilon_1 + \kappa(A - \sigma I)\epsilon_2}{1 - \kappa(A - \sigma I)\epsilon_2}\|v_j\|.$$

Proof We have

$$\begin{aligned}\|r_j\| &\leq \|A - \sigma I\|\|(A - \sigma I)^{-1}(v_j + r_j)\|\epsilon_2 + \|v_j\|\epsilon_1 \\ &\leq \kappa(A - \sigma I)\|v_j + r_j\|\epsilon_2 + \|v_j\|\epsilon_1 \\ &\leq \kappa(A - \sigma I)(\|v_j\| + \|r_j\|)\epsilon_2 + \|v_j\|\epsilon_1.\end{aligned}$$

Reordering gives the result. □

The following result yields a family of new residual bounds independent of $\|v_j\|$.

Proposition 3 *Let* $(A - \sigma I)^{-1}(v_j + r_j) = w_j$ *and assume* (6) *holds. If*

$$0 < \frac{\epsilon_1 + \kappa(A - \sigma I)\epsilon_2}{1 - \kappa(A - \sigma I)\epsilon_2} \leq \gamma < 1, \tag{7}$$

then

$$\|r_j\| \le \left(\epsilon_2 + \frac{\epsilon_1}{1-\gamma}\right) \|A - \sigma I\| \|w_j\|.$$

Proof From (6) we have

$$\|r_j\| \le \left(\epsilon_2 + \epsilon_1 \frac{\|v_j\|}{\|A - \sigma I\| \|w_j\|}\right) \|A - \sigma I\| \|w_j\|.$$

Thus we need to show $\|v_j\|/(\|A - \sigma I\| \|w_j\|) \le 1/(1-\gamma)$. We have

$$\frac{\|v_j\|}{\|A - \sigma I\| \|w_j\|} = \frac{\|v_j\|}{\|A - \sigma I\| \|(A - \sigma I)^{-1}(v_j + r_j)\|} \le \frac{\|v_j\|}{\|v_j + r_j\|},$$

and from the reverse triangle inequality,

$$\frac{\|v_j\|}{\|v_j + r_j\|} \le \frac{\|v_j\|}{|\|v_j\| - \|r_j\||}.$$

Now, by Lemma 2 and assumption (7), we have

$$\|r_j\| \le \frac{\epsilon_1 + \kappa(A - \sigma I)\epsilon_2}{1 - \kappa(A - \sigma I)\epsilon_2} \|v_j\| \le \gamma \|v_j\|.$$

Putting everything together yields

$$\frac{\|v_j\|}{\|A - \sigma I\| \|w_j\|} \le \frac{\|v_j\|}{|\|v_j\| - \|r_j\||} \le \frac{1}{1-\gamma}.$$

□

In particular, if $\kappa(A - \sigma I) \le (1 - 2\epsilon_1)/(3\epsilon_2)$, then we have $\kappa(A - \sigma I)\epsilon_2 < 1$ and can take $\gamma = 1/2$ in Proposition 3, to obtain

$$\|r_j\| \le (2\epsilon_1 + \epsilon_2) \|A - \sigma I\| \|w_j\|. \tag{8}$$

This is the same bound as we get from (6) if we replace (ϵ_1, ϵ_2) with $(0, 2\epsilon_1 + \epsilon_2)$. In particular, if the linear systems are solved in a backward stable manner so that (4) holds, and $\kappa(A - \sigma I) \le (1 - 2\epsilon_{\text{bw}})/(3\epsilon_{\text{bw}})$, then (8) holds with $2\epsilon_1 + \epsilon_2 = 3\epsilon_{\text{bw}}$.

3 Errors from orthonormalization

In this section we are concerned with the orthonormalization error

$$g_j = w_j - V_{j+1} h_{1:j+1,j}.$$

Up to signs, this error can be viewed as the backward error in the $(j+1)$st column of a perturbed QR factorization

$$[v_1\ w_1\ w_2\ \cdots\ w_k] = V_{k+1}[e_1\ \underline{H_k}] + [0\ g_1\ g_2\ \cdots\ g_k]. \tag{9}$$

Thus, we are interested in *columnwise* backward error bounds for QR factorizations. The next theorem shows how such bounds can be obtained from normwise backward error bounds given in the 2-norm or the Frobenius norm. It applies to floating point algorithms $\mathtt{qr}(\cdot)$ that are unaffected by power-of-two column scalings, in the sense that if $[Q, R] = \mathtt{qr}(A)$, then $[Q, RD] = \mathtt{qr}(AD)$ for any $D = \mathrm{diag}(d_1, d_2, \ldots, d_k)$ where the d_i are powers of 2. Barring underflow and overflow, this covers commonly used QR algorithms such as classical and modified Gram-Schmidt with and without (possibly partial) reorthogonalization, Householder QR and Givens QR.

Theorem 4 *Let* $\mathtt{qr}(\mathrm{A})$ *denote an algorithm that computes an approximate QR factorization of an* $n \times k$ *matrix* A *in floating point arithmetic. Suppose further that* $[Q, RD] = \mathtt{qr}(AD)$ *for any* $D = \mathrm{diag}(d_1, d_2, \ldots, d_k)$ *where the* d_i *are powers of 2. If* Q *and* R *denote the computed factors,* $\Delta A = A - QR$ *and* $\|\Delta A\|_* \le \gamma \|A\|_* u$, *where* $\|\cdot\|_*$ *denotes the 2-norm or the Frobenius norm, then* $\|\Delta a_i\| \le 2\gamma\sqrt{k}\|a_i\|u$ *for* $i = 1\!:\!k$.

Proof For $i = 1\!:\!k$, we define

$$d_i = 2^{-\lfloor \log_2 \|a_i\| \rfloor},$$

so $1 \le \|a_i\| d_i < 2$. Since ΔAD is the backward error from $\mathtt{qr}(AD)$ we have

$$\begin{aligned} d_i \|\Delta a_i\| &= \|\Delta AD e_i\| \le \|\Delta AD\|_* \\ &\le \gamma \|AD\|_* u < 2\gamma\sqrt{k}\|ADe_i\|u = d_i 2\gamma\sqrt{k}\|a_i\|u, \end{aligned}$$

for $i = 1 : k$, from which the theorem follows. □

The constant γ in Theorem 4 is obviously algorithm dependent and many bounds exist in the literature. Some of them contain both n and k [23], and others only k [1, 5], [12, Theorem 19.13]. In [12, p. 361] a columnwise bound depending on n and k is given. For Krylov methods we usually have $n \gg k$, so bounds independent from n should certainly be favored. We shall assume that

$$\|g_j\| \le \eta(n, k)\|w_j\|u, \tag{10}$$

holds for some function $\eta(n, k)$.

3.1 Columnwise backward errors for modified Gram–Schmidt

Our next theorem shows that for MGS, with and without one round of reorthogonalization, η in (10) does not depend on n and is given by

$$\eta(n,k) = \zeta k,$$

where ζ is a modest constant. We need the following forward error result for _axpy operations.

Lemma 5 *Let α be a scalar and x and y vectors. If*

$$s = \text{float}(\alpha x + y) - (\alpha x + y) \quad \textit{then} \quad \|s\| \leq 2(\|\alpha x\| + \|y\|)u.$$

Proof The ith component of $\alpha x + y$ can be viewed as the inner product $[x_i \ y_i][\alpha \ 1]^T$. Thus the componentwise forward error is bounded by $|s| \leq 2u(|\alpha x| + |y|)$ [14]. We get

$$\|s\| \leq \|2u(|\alpha x| + |y|)\| \leq 2(\|\alpha x\| + \|y\|)u.$$

□

The next theorem gives columnwise backward error bounds for MGS with and without one round of reorthogonalization.

Theorem 6 *Let Q and R denote the computed factors in a QR decomposition of an $n \times k$ matrix A, which was obtained by a floating point implementation of modified Gram-Schmidt with or without one round of reorthogonalization. Assume*

(i) $\|q_j\| = 1$ *for* $j = 1:k$, *and*
(ii) $(1 + (n+3)u)^k < 1 + \delta$ *for some* $\delta > 0$.

Then there exists a ΔA such that $A + \Delta A = QR$ with $\|\Delta a_j\| \leq cj\|a_j\|u$, where $c = 4(1+\delta)$ if no reorthogonalization was done and $c = 10(1+\delta)^2$ if one round of reorthogonalization was done.

Let us pause for a while and discuss the assumptions before we proceed with the proof. Assumption (i) is imposed to keep our analysis cleaner; it does not affect our final bounds in any significant way. Assumption (ii) is needed for the following reason: if we compute $y = \text{float}\left(x - q_j(q_j^H x)\right)$ for some $1 \leq j \leq k$, then, assuming (i), the quantity $1 + (n+3)u = 1 + \|q_j\|^2(n+3)u$ is an upper bound for $\|y\|/\|x\|$ [12, Lemma 3.9]. Thus, (ii) guarantees that we can apply a sequence of k elementary "floating point" projections of the form $I - q_i q_i^H$ to any vector x, and the resulting vector will be bounded in norm by $(1+\delta)\|x\|$.

Proof of Theorem 6 Let $R^{(1)}$ and $R^{(2)}$ denote the *strictly* upper triangular matrices containing the orthogonalization coefficients corresponding to the first and second round of orthogonalization, respectively. We define $R^{(2)} \equiv 0$, if no reorthogonalization

is done. Assume for a while that $R^{(1)}$ and $R^{(2)}$ are given, and suppose we want to compute

$$a_j - \sum_{i=1}^{j-1} r_{ij}^{(1)} q_i - \sum_{i=1}^{j-1} r_{ij}^{(2)} q_i.$$

This can be viewed as $2(j-1)$ _axpy operations. We define $a_j^{(0)} = a_j$ and

$$a_j^{(i)} = \begin{cases} \text{float}(a_j^{(i-1)} - r_{ij}^{(1)} q_i) & \text{for} \quad i = 1 : j-1, \\ \text{float}(a_j^{(i-1)} - r_{(i-j+1)j}^{(2)} q_{i-j+1}) & \text{for} \quad i = j : 2(j-1). \end{cases}$$

Using Lemma 5 yields

$$a_j^{(i)} = \begin{cases} a_j^{(i-1)} - r_{ij}^{(1)} q_i + s_i & \text{for} \quad i = 1 : j-1, \\ a_j^{(i-1)} - r_{(i-j+1)j}^{(2)} q_{i-j+1} + s_i & \text{for} \quad i = j : 2(j-1), \end{cases}$$

where

$$\|s_i\| \leq \begin{cases} 2(\|r_{ij}^{(1)} q_i\| + \|a_j^{(i-1)}\|)u & \text{for} \quad i = 1 : j-1, \\ 2(\|r_{(i-j+1)j}^{(2)} q_{i-j+1}\| + \|a_j^{(i-1)}\|)u & \text{for} \quad i = j : 2(j-1). \end{cases}$$

Now, $a_j^{(i-1)}$ is also the result of applying $i-1$ elementary floating point projections to a_j, so the discussion prior to the proof gives $\|a_j^{(i-1)}\| < (1+(n+3)u)^{i-1}\|a_j\|$. Further, from (ii) we have $(1+nu)\|a_j^{(i-1)}\| < (1+\delta)\|a_j\|$ for $i = 1 : j-1$ and $(1+nu)\|a_j^{(i-1)}\| < (1+\delta)^2\|a_j\|$ for $i = j : 2(j-1)$. The forward error of a computed inner product $\text{float}(x^H y)$, where x and y are of length n, is bounded by $nu\|x\|\|y\|$ [14]. Thus,

$$\left|r_{ij}^{(1)}\right| \leq \left|\text{float}\left(q_i^H a_j^{(i-1)}\right)\right| \leq \left|q_i^H a_j^{(i-1)}\right| + nu\left\|a_j^{(i-1)}\right\| < (1+\delta)\|a_j\|$$

and, similarly, $\left|r_{ij}^{(2)}\right| < (1+\delta)^2\|a_j\|$. Thus, s_i is bounded by

$$\|s_i\| \leq \begin{cases} 4(1+\delta)\|a_j\|u & \text{for} \quad i = 1 : j-1, \\ 4(1+\delta)^2\|a_j\|u & \text{for} \quad i = j : 2(j-1). \end{cases}$$

We have

$$a_j - \sum_{i=1}^{j-1} r_{ij}^{(1)} q_i - \sum_{i=1}^{j-1} r_{ij}^{(2)} q_i = a_j^{(2(j-1))} - \sum_{i=1}^{2(j-1)} s_i.$$

If we define $d_i = \text{float}\left(\|a_j^{(2(j-1))}\|\right)$ and $q_j = \text{float}\left(a_j^{(2(j-1))}/d_j\right)$ and note that

$$a_j^{(2(j-1))} = q_j d_j + f_j \quad \text{with} \quad \|f_j\| \leq \left\|a_j^{(2(j-1))}\right\| u < (1+\delta)^2 \|a_j\| u,$$

then we get

$$a_j - \sum_{i=1}^{j-1} (r_{ij}^{(1)} + r_{ij}^{(2)}) q_i - d_j q_j = f_j - \sum_{i=1}^{2(j-1)} s_i.$$

Finally, defining $R = \text{float}(R^{(1)} + R^{(2)}) + \text{diag}(d_1, d_2, \ldots, d_k)$ yields

$$\Delta a_j := a_j - \sum_{i=1}^{j} r_{ij} q_i = f_j - \sum_{i=1}^{2(j-1)} s_i - \sum_{i=1}^{j-1} \Delta r_{ij} q_i,$$

where

$$\Delta r_{ij} = r_{ij}^{(1)} + r_{ij}^{(2)} - r_{ij}, \quad \text{so} \quad |\Delta r_{ij}| \leq |r_{ij}^{(1)} + r_{ij}^{(2)}| u < 2(1+\delta)^2 \|a_j\| u.$$

Using the above bounds for f_j, the s_i and the Δr_{ij} gives $\|\Delta a_j\| < 10(1+\delta)^2 j \|a_j\| u$. If no reorthogonalization was done, then we have $s_i = 0$ for $i = j\!:\!2(j-1)$, and $\Delta r_{ij} = 0$, $\|f_j\| \leq (1+\delta)\|a_j\| u$ for all j. Taking this into account yields $\|\Delta a_j\| < 4(1+\delta) j \|a_j\| u$. □

Remark 1 Suppose the perturbed QR factorization (9) was computed using MGS. Then, by taking $\delta = 1/10$ and assuming that the conditions of Theorem 6 hold, we get that $\eta(n,k)$ in (10) is bounded by $\eta(n,k) \leq 5k$ if standard MGS is used, and $\eta(n,k) \leq 13k$ if MGS with one round of reorthogonalization is used. We point out that these bounds should not be interpreted as saying that standard MGS should be favored over MGS with reorthogonalization. On the contrary, as we will see in the next section, retaining a well-conditioned basis (which is the effect of reorthogonalization) is of great importance to the shift-and-invert Arnoldi algorithm.

4 Backward error bounds for the shift-and-invert Arnoldi recurrence

Recall the perturbed Krylov recurrence

$$(A - \sigma I)^{-1}(V_k + F_k) = V_{j+1} \underline{H}_k, \tag{11}$$

where $F_k = [f_1 \ f_2 \ \cdots \ f_k]$ and f_j, for $j = 1\!:\!k$, is defined by $f_j = r_j - (A - \sigma I) g_j$. We discussed in Sects. 2 and 3 how to bound r_j and g_j, respectively. By using these bounds, we can now easily bound F_k. Assuming (6) and (10) yields

$$\|f_j\| \leq \|v_j\| \epsilon_1 + \|A - \sigma I\| \|w_j\| (\epsilon_2 + \eta(n, j) u). \tag{12}$$

Further, from (9) we see that

$$\|w_j\| = \|V_{j+1}h_{1:j+1,j} + g_j\| \leq \|V_{j+1}\|\|h_{1:j+1,j}\| + \eta(n, j)\|w_j\|u,$$

which in turn implies

$$\|w_j\| \leq \frac{\|V_{j+1}\|\|h_{1:j+1,j}\|}{1 - \eta(n, j)u},$$

assuming that $\eta(n, j)u < 1$. We get

$$\|f_j\| \leq \|v_j\|\epsilon_1 + \|A - \sigma I\|\|V_{j+1}\|\|h_{1:j+1,j}\|c_{jn}(\epsilon_2)$$

and further (assuming that $\eta(n, k)$ is monotonically increasing in k)

$$\|F_k\| \leq \sqrt{k}\|V_k\|\epsilon_1 + \sqrt{k}\|A - \sigma I\|\|V_{k+1}\|\|\underline{H}_k\|c_{kn}(\epsilon_2), \tag{13}$$

where

$$c_{kn}(\epsilon_2) := \frac{\epsilon_2 + \eta(n, k)u}{1 - \eta(n, k)u} \tag{14}$$

should be thought of as a tiny factor.

Similarly, if we assume the bound (8) instead of (6), we get

$$\|F_k\| \leq \sqrt{k}\|A - \sigma I\|\|V_{k+1}\|\|\underline{H}_k\|c_{kn}(2\epsilon_1 + \epsilon_2). \tag{15}$$

This is the same bound we get from (13) if we replace (ϵ_1, ϵ_2) by $(0, 2\epsilon_1 + \epsilon_2)$.

Having established (13) and (15), we are now ready to reshuffle Eq. (11) in order to derive backward error bounds for the shift-and-invert Krylov recurrence. We will derive perturbed recurrences of the form

$$V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{H}_k. \tag{16}$$

If we look at this from a backward error perspective, (16) means that we have taken k steps, without errors, of a shift-and-invert Krylov algorithm applied to a perturbed pencil, and all linear systems that occurred in the process must have been consistent. However, in order to rewrite (16) as

$$(A + \Delta A - \sigma I)^{-1}V_k = V_{k+1}\underline{H}_k,$$

we need to ensure that $A + \Delta A - \sigma I$ is invertible. We need the following lemma to solve this technicality.

Lemma 7 *Let A and V be matrices of size $n \times n$ and $n \times k$ respectively, such that* $\operatorname{rank} AV = k$. *Then for any $\epsilon > 0$, there exists a matrix X with $\|X\| < \epsilon$ such that $A + X$ is nonsingular and $XV = 0$. Furthermore, if A is Hermitian, then we may take X to be Hermitian.*

Proof Find a unitary matrix Q such that

$$Q^H V = \begin{bmatrix} 0 \\ V_2 \end{bmatrix} \tag{17}$$

for some $k \times k$ matrix V_2, and define $AQ = [A_1 \ A_2]$ where A_2 is of size $n \times k$. From $\text{rank} AV = k$, it follows A_2 has rank k. Define Y so its columns span the orthogonal complement to range of A_2, and set $Z = [Y - A_1 \ 0]$. We have that $A + ZQ^H = [Y \ A_2]Q^H$ is nonsingular and $ZQ^H V = 0$. In particular, this means that the pencil $A+\lambda ZQ^H$ is regular. If λ is any value outside the spectrum of the pencil such that $|\lambda| < \epsilon/\|Z\|$, then $X = \lambda ZQ^H$ satisfies the conditions of the theorem.

For the second part, suppose A is Hermitian and Q is such that (17) holds. Write

$$Q^H AQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^H & A_{22} \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} \omega I - A_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad \omega > 0,$$

where A_{11} is of size $(n-k)\times(n-k)$. We have that QWQ^H is Hermitian, $QWQ^H V = 0$, and

$$Q(Q^H AQ + W)Q^H = A + QWQ^H.$$

Thus, for the same reason as above, it is enough to find one $\omega > 0$ such that $Q^H AQ + W$ is nonsingular. Let

$$A_{22} = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} U^H$$

be a spectral decomposition where D is of full rank, and define $[B_1 \ B_2] = A_{12}U$ such that B_1 has as many columns as D. We have that $Q^H AQ + W$ is nonsingular if and only if

$$\begin{bmatrix} \omega I & B_1 & \omega B_2 \\ B_1^H & D & 0 \\ \omega B_2^H & 0 & 0 \end{bmatrix}$$

is nonsingular. Further, since $[A_{12}^T \ A_{22}^T]^T$ is of full rank, and

$$\begin{bmatrix} B_1 & B_2 \\ D & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & U^H \end{bmatrix} \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} U,$$

it follows that B_2 is also of full rank. We have

$$\begin{bmatrix} \omega I & B_1 & \omega B_2 \\ B_1^H & D & 0 \\ \omega B_2^H & 0 & 0 \end{bmatrix} \begin{bmatrix} I & -\omega^{-1} B_1 & -\omega^{-1} B_2 \\ 0 & I & 0 \\ 0 & 0 & \omega^{-1} I \end{bmatrix}$$
$$= \begin{bmatrix} \omega I & 0 & 0 \\ B_1^H & D - \omega^{-1} B_1^H B_1 & -\omega^{-1} B_1^H B_2 \\ \omega B_2^H & -B_2^H B_1 & -B_2^H B_2 \end{bmatrix},$$

which is easily seen to be nonsingular for large enough values of ω. □

If we use the bound on F_k shown in (13), then we can deduce the following theorem.

Theorem 8 *Let $(A - \sigma I)^{-1}(V_k + F_k) = V_{k+1}\underline{H}_k$ be of full rank and assume F_k is bounded as in (13) and $\sqrt{k}\kappa(V_k)\epsilon_1 < 1$. Then there is a ΔA of rank at most k such that*

$$V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{H}_k,$$

and

$$\|\Delta A\| \leq \sqrt{k}\|A - \sigma I\| \frac{\kappa(V_k)\epsilon_1 + \kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(\epsilon_2)}{1 - \sqrt{k}\kappa(V_k)\epsilon_1},$$

where $c_{kn}(\epsilon_2)$ is given by (14).

Proof From $V_k + F_k = (A - \sigma I)V_{k+1}\underline{H}_k$ and $V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{H}_k$ we see that any eligible ΔA has to satisfy $\Delta A V_{k+1}\underline{H}_k = -F_k$. We choose $\Delta A = -F_k(V_{k+1}\underline{H}_k)^\dagger$ (which is of rank at most k) which implies $\|\Delta A\| \leq \|F_k\|/\sigma_{\min}(V_{k+1}\underline{H}_k)$. Substituting $\|F_k\|$ by the upper bound given in (13) yields

$$\begin{aligned} \|\Delta A\| &\leq \frac{\sqrt{k}\|V_k\|\epsilon_1 + \sqrt{k}\|A - \sigma I\|\|V_{k+1}\|\|\underline{H}_k\|c_{kn}(\epsilon_2)}{\sigma_{\min}(V_{k+1}\underline{H}_k)} \\ &\leq \frac{\sqrt{k}\|V_k\|\epsilon_1}{\sigma_{\min}(V_{k+1}\underline{H}_k)} + \sqrt{k}\|A - \sigma I\|\kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(\epsilon_2). \end{aligned}$$

For the denominator we get

$$\begin{aligned} \sigma_{\min}(V_{k+1}\underline{H}_k) &\geq \sigma_{\min}\big((A + \Delta A - \sigma I)V_{k+1}\underline{H}_k\big)/\|A + \Delta A - \sigma I\| \\ &\geq \sigma_{\min}(V_k)/(\|A - \sigma I\| + \|\Delta A\|), \end{aligned}$$

where we used $\sigma_{\min}(XY) \leq \|X\|\sigma_{\min}(Y)$ which holds for any matrices X, Y. Thus

$$\|\Delta A\| \leq \frac{\sqrt{k}\|V_k\|(\|A - \sigma I\| + \|\Delta A\|)\epsilon_1}{\sigma_{\min}(V_k)} + \sqrt{k}\|A - \sigma I\|\kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(\epsilon_2)$$

which can be reordered to the claimed bound. □

If the linear systems are solved up to a normwise backward error ϵ_{bw}, and (8) and (15) hold for $2\epsilon_1 + \epsilon_2 = 3\epsilon_{\text{bw}}$, then we get the following corollary.

Corollary 9 *Let $(A - \sigma I)^{-1}(V_k + F_k) = V_{k+1}\underline{H}_k$ be of full rank and assume F_k is bounded as in (15) with $2\epsilon_1 + \epsilon_2 = 3\epsilon_{\text{bw}}$. Then there is a ΔA of rank at most k such that*

$$V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{H}_k,$$

and

$$\|\Delta A\| \leq \sqrt{k}\|A - \sigma I\|\kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(3\epsilon_{\text{bw}}),$$

where $c_{kn}(\cdot)$ is given by (14).

A few remarks are in order.

Remark 2 If $A + \Delta A - \sigma I$ in Theorem 8 and Corollary 9 is singular, then we can invoke Lemma 7 with $V = V_{k+1}\underline{H}_k$ to obtain a backward error $\Delta\widehat{A}$, arbitrarily close to ΔA, such that $(A + \Delta\widehat{A} - \sigma I)^{-1}V_k = V_{k+1}\underline{H}_k$. The new backward error $\Delta\widehat{A}$ will in general have rank greater than k, but its numerical rank is still bounded by k. Here the definition of numerical rank can be arbitrarily strict, in the sense that we may define the numerical rank as the number of singular values that greater than $\epsilon > 0$, for an arbitrarily small ϵ.

Remark 3 If the orthonormalization is done properly, using, for instance, MGS with reorthogonalization, then $\kappa(V_{k+1}) \approx 1$. In this case we can ignore the factors $\kappa(V_{k+1})$ and $\kappa(V_k)$ when evaluating the bounds in Theorem 8 and Corollary 9. In particular this means that the bounds can be estimated cheaply as long as $\|A - \sigma I\|$ (or a good estimate of it) is known.

Remark 4 For the standard eigenvalue problem, shifts are used to find interior eigenvalues, so any sensible shift satisfies $|\sigma| \leq \|A\|$. Thus, we have $\|A - \sigma I\| \leq 2\|A\|$ in practice.

Remark 5 In view of [6], we note that our bounds do not contain the loss-of-orthonormality term $\|V_{k+1}^H V_{k+1} - I\|$. Instead we saw that the condition number of the computed basis V_{k+1} plays a role in the bounds of the backward error. We note, however, that a small value of $\|V_{k+1}^H V_{k+1} - I\|$ implies that V_{k+1} is well-conditioned:

$$\|V_{k+1}^H V_{k+1} - I\| < \epsilon < 1 \quad \Rightarrow \quad \kappa(V_{k+1}) < \sqrt{\frac{1+\epsilon}{1-\epsilon}}.$$

The next example shows how Theorem 8 can be used to derive a simple a posteriori backward error bound.

Example 1 Suppose a matrix A and a shift σ with $|\sigma| < \|A\|$ are given, and suppose we perform k steps of the shift-and-invert Arnoldi algorithm. To solve the linear systems

we use an iterative method that employs (5) as stopping condition, that is, the linear systems are considered "solved" when the residuals are less than some tolerance ϵ_{tol} (we ignore the norm of the right hand side since it is approximately one). We use a rather crude tolerance so $\epsilon_{\text{tol}} \gg u$. For the orthogonalization we use MGS with one round of reorthogonalization so $c_{kn}(0) \lesssim 13ku$ (cf. Remark 1). If

$$\epsilon_{\text{tol}} \geq \kappa(\underline{H}_k)c_{kn}(0), \tag{18}$$

then Theorem 8, with $\epsilon_1 = \epsilon_{\text{tol}}$ and $\epsilon_2 = 0$, and the following remarks, yield that the computed quantities satisfy

$$(A + \Delta A - \sigma I)^{-1} V_k = V_{k+1}\underline{H}_k,$$

where

$$\|\Delta A\| \leq \frac{4\sqrt{k}\kappa(V_{k+1})\epsilon_{\text{tol}}}{1 - \sqrt{k}\kappa(V_k)\epsilon_{\text{tol}}}\|A\|. \tag{19}$$

Here we have used the fact that $\kappa(V_{k+1}) \geq \kappa(V_k)$. Since MGS with reorthogonalization was employed, we expect $\kappa(V_{k+1})$ to be close to one. Thus, (19) tells us that the relative backward error $\|\Delta A\|/\|A\|$ is a modest multiple of the tolerance we used to solve the linear systems. *So, in this setting the shift-and-invert Arnoldi algorithm is backward stable.*

We end this section with a numerical experiment. We consider two matrices of order $n = 1000$ and associated shifts. The first matrix is the symmetric tridiagonal matrix

$$A_1 = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{bmatrix},$$

and the associated shift is $\sigma_1 = -2$. It is well-known that the eigenvalues of A_1 are given by $-2 + 2\cos(\pi k/(n+1))$, for $k = 1:n$, so $A_1 - \sigma_1 I$ is indeed invertible. The second matrix is the nonnormal matrix

$$A_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & & & \\ -1 & 1 & 1 & 1 & 1 & & \\ & -1 & 1 & 1 & 1 & 1 & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & -1 & 1 & 1 & 1 & 1 \\ & & & & -1 & 1 & 1 & 1 \\ & & & & & -1 & 1 & 1 \\ & & & & & & -1 & 1 \end{bmatrix},$$

also known as the Grcar matrix [11]. The associated shift was chosen to be $\sigma_2 = 1$. It is an easy exercise to show that $A_2 - \sigma_2 I$ is invertible.

We implemented the shift-and-invert Arnoldi algorithm in MATLAB R2013a. For orthonormalization we used MGS with one round of reorthogonalization. The matrices were stored in sparse format, and the linear systems were solved using MATLAB's "backslash" and `lu` routines. We took $k = 30$ steps with the starting vector $[1, 1, \ldots, 1]^T$, and in each iteration we computed the backward error shown in (3), where the residual was evaluated in extended precision (32 digits) and then rounded to double precision. We did this using the `vpa` function from the Symbolic Math Toolbox. We also computed the errors $F_k = V_k - (A_i - \sigma_i I)V_{k+1}\underline{H}_k$, $i = 1, 2$, in extended precision and rounded the result to double precision. For each $j = 1{:}k$ and $i = 1, 2$, we computed

$$\mathcal{B}\left(\left\|\Delta A_i^{(j)}\right\|\right) := \sqrt{j}\|A_i - \sigma_i I\|\kappa(\underline{H}_j)c_{jn}(3\epsilon_{\mathrm{bw}}),$$

where ϵ_{bw} was set to be the largest backward error of the linear systems that was encountered in the algorithm, and $c_{jn}(3\epsilon_{\mathrm{bw}}) := (3\epsilon_{\mathrm{bw}} + 13ju)/(1 - 13ju)$ (cf. Remark 1). As is mentioned in Remark 3, the above quantity is a good estimate of the bound in Corollary 9. We also evaluated the expression for the backward errors, $\Delta A_i^{(j)} = -F_k(V_{k+1}\underline{H}_k)^\dagger$, $i = 1, 2$, given in the proof of Theorem 8, and computed their norms. We did this using the MATLAB routines `pinv` (for the Moore–Penrose pseudo-inverse) and `norm`. The quantities $\mathcal{B}(\|\Delta A_i^{(j)}\|)$ and $\|\Delta A_i^{(j)}\|$ are shown in Fig. 1 for $j = 1{:}30$, $i = 1, 2$.

Our experiment seems to be unaffected by the nonnormality of A_2. Moreover, even though the (estimated) upper bounds $\mathcal{B}(\|\Delta A_i^{(j)}\|)$, $i = 1, 2$, can be seen to be rather pessimistic, they do show that the backward errors are less than $\sqrt{u}$. In other words, for both matrices, the upper bounds show that the computation is backward stable up to single precision.

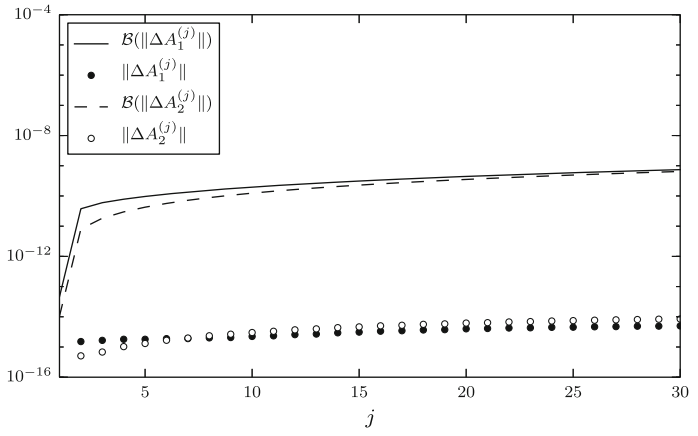

Fig. 1 Computed backward errors and associated bound

5 Further topics

5.1 Implicit restarting

The bounds in Theorem 8 and Corollary 9 contain the factor $\kappa(\underline{H}_k)$, so if $\kappa(\underline{H}_k) \gg 1$ we cannot guarantee a small backward error. If we recall how Arnoldi locates eigenvalues [25, pp. 257–265], we have, unfortunately, reason to suspect that this is the case. Since Arnoldi does not target the largest eigenvalues, but *any* isolated eigenvalue cluster, $H_k := [I_k \ 0]\underline{H}_k$ is likely to have both large and small eigenvalues, which suggests that H_k may be ill-conditioned. We will now show that the situation can be much better than expected if we restrict our attention to the largest eigenvalues of H_k, that is, the ones corresponding to eigenvalues of A closest to the shift σ. The idea is to do an implicit (thick) restart [24], and purge the small eigenvalues of H_k. Since small eigenvalues of H_k correspond to eigenvalues of A far from the shift σ, it is reasonable to assume they are of less interest. Suppose

$$(A - \sigma I)^{-1}(V_k + F_k) = V_{k+1}\underline{H}_k$$

and consider a Schur form $H_k = QTQ^H$ such that t_{ii}, $i = \ell + 1 : k$, are the small eigenvalues to be purged. We have

$$(A - \sigma I)^{-1}(U_k + F_k Q) = [U_k \ v_{k+1}] \begin{bmatrix} T \\ h_{k+1,k} e_k^T Q \end{bmatrix},$$

where $U_k = V_k Q$. Throwing away the last $k - \ell$ columns yields

$$(A - \sigma I)^{-1}(U_\ell + F_k Q_\ell) = [U_\ell \ v_{k+1}] \begin{bmatrix} T_\ell \\ h_{k+1,k} e_k^T Q_\ell \end{bmatrix},$$

where $Q_\ell = Q(:, 1:\ell)$, $U_\ell = U(:, 1:\ell)$ and $T_\ell = T(1:\ell, 1:\ell)$. Defining $u_{\ell+1} = v_{k+1}$,

$$\underline{T}_\ell = \begin{bmatrix} T_\ell \\ h_{k+1,k} e_k^T Q_\ell \end{bmatrix},$$

and $E_\ell = F_k Q_\ell$, results in a compact recurrence

$$(A - \sigma I)^{-1}(U_\ell + E_\ell) = U_{\ell+1}\underline{T}_\ell, \tag{20}$$

where $\|E_\ell\| \leq \|F_k\|$. Note that our bound on E_ℓ depends on k and not ℓ. We can now repeat the proof of Theorem 8, and use the bounds $\|E_\ell\| \leq \|F_k\|$ and $\sigma_{\min}(U_{\ell+1}) \geq \sigma_{\min}(V_{k+1})$, and the recurrence (20) instead of the one assumed in the theorem. We get

$$U_\ell = (A + \Delta A - \sigma I)U_{\ell+1}\underline{T}_\ell,$$

where

$$\|\Delta A\| \leq \|A - \sigma I\| \frac{\sqrt{k}\kappa(V_k)\epsilon_1 + \sqrt{k}\kappa(V_{k+1})c_{kn}(\epsilon_2)\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell)}{1 - \sqrt{k}\kappa(V_k)\epsilon_1}. \tag{21}$$

Comparing this to the bound in Theorem 8 we see that $\kappa(\underline{H}_k)$ has been replaced by $\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell)$. Further, it holds that

$$\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell) \leq \|\underline{H}_k\|/\sigma_{\min}\left(\begin{bmatrix} T \\ h_{k+1,k}e_k^T Q \end{bmatrix}\right) = \kappa(\underline{H}_k).$$

It follows that if $\underline{H}_k$ is ill-conditioned due to the small eigenvalues we purged, then $\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell) \ll \kappa(\underline{H}_k)$ and (21) shows that the upper bound for the backward error corresponding to the part of the spectrum we care about is much smaller than the upper bound for the general backward error.

5.2 Hermitian backward errors

We now restrict the scope to the Hermitian matrix eigenvalue problem, that is, when $A = A^H$ and σ is real. Let us mention that we still consider the shift-and-invert Arnoldi algorithm, as it is shown in Algorithm 1, and *not* the shift-and-invert Lanczos algorithm with a three-term recurrence. In the Hermitian case, Algorithm 1 is also known as the shift-and-invert Lanczos algorithm with full orthogonalization, and it is used in, e.g., ARPACK [17, routine `ssaitr.f`] and MATLAB's `eigs` command.

Is it, for a Hermitian A, possible to find a Hermitian backward error ΔA? We have seen in the proof of Theorem 8 that ΔA has to satisfy $\Delta A V_{k+1}\underline{H}_k = -F_k$. Unfortunately the following lemma rules out existence of such a Hermitian ΔA in general.

Lemma 10 *Let $X \in \mathbb{C}^{n\times k}$ and $F \in \mathbb{C}^{n\times k}$. Then there exists a Hermitian E with $EX = F$ if and only if $X^H F$ is Hermitian and $FX^\dagger X = F$. In that case, there is such an E with* $\operatorname{rank}(E) \leq 2k$ *and $\|E\|_* \leq 2\|F\|_*/\sigma_{\min}(X)$ where $\|\cdot\|_*$ denotes the 2-norm or the Frobenius norm.*

Proof The proof is simple and, for $k = 1$, is contained in [18]. We give it for completeness. Let E be any matrix such that $EX = F$. This implies $EXX^\dagger X = FX^\dagger X$ and (using $XX^\dagger X = X$) $EX = FX^\dagger X$, contradicting $EX = F$ if $F \neq FX^\dagger X$. Thus $F = FX^\dagger X$ is necessary for the existence of an E with $EX = F$. Now, if E is Hermitian, then so is $X^H EX = X^H F$. Hence, if $X^H F$ is not Hermitian, then there is no Hermitian E with $EX = F$.

On the other hand, if $X^H F$ is Hermitian and $F = FX^\dagger X$, then

$$E := FX^\dagger + (FX^\dagger)^H - X^{\dagger H}F^H XX^\dagger = FX^\dagger + (FX^\dagger)^H(I - XX^\dagger)$$

is also Hermitian. Furthermore, $\operatorname{rank}(E) \leq 2k$, $EX = F$, and (using that $I - XX^\dagger$ is an orthogonal projector)

$$\|E\|_* \leq 2\|FX^\dagger\|_* \leq 2\|F\|_*\|X^\dagger\|_2 = 2\|F\|_*/\sigma_{\min}(X).$$

□

The next result shows that one still gets a Hermitian backward error if one replaces the Hessenberg matrix $\underline{H}_k$ by some other $(k+1) \times k$ matrix $\underline{G}_k$. Before we state the theorem, we should clarify what we mean by "backward error" in this case. If we replace $\underline{H}_k$ by something else, we cannot say that the computed quantities (V_{k+1} and $\underline{H}_k$) satisfy a an exact Krylov recurrence of a perturbed input matrix. We can, however, still say that the *computed subspace* is a Krylov subspace of a perturbed Hermitian input matrix. We refer to this Hermitian perturbation as the backward error.

Theorem 11 *Let A be Hermitian and $(A - \sigma I)^{-1}(V_k + F_k) = V_{k+1}\underline{H}_k$. Suppose it holds for $\underline{G}_k \in \mathbb{C}^{(k+1)\times k}$ that $V_k^H V_{k+1}\underline{G}_k$ is Hermitian and $V_{k+1}\underline{G}_k$ is of full rank. Then there is a Hermitian ΔA of rank at most $2k$ such that*

$$V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{G}_k,$$

and

$$\|\Delta A\| \leq 2\frac{\|(A - \sigma I)\|\|V_{k+1}\|\|\underline{H}_k - \underline{G}_k\| + \|F_k\|}{\sigma_{\min}(V_{k+1}\underline{G}_k)}.$$

Proof From $V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{G}_k$ and

$$V_k + F_k = (A - \sigma I)V_{k+1}\underline{H}_k = (A - \sigma I)V_{k+1}\underline{G}_k + (A - \sigma I)V_{k+1}(\underline{H}_k - \underline{G}_k)$$

we see that any eligible ΔA has to satisfy

$$\Delta A V_{k+1}\underline{G}_k = (A - \sigma I)V_{k+1}(\underline{H}_k - \underline{G}_k) - F_k = V_k - (A - \sigma I)V_{k+1}\underline{G}_k.$$

Since it is assumed that $V_{k+1}\underline{G}_k$ is of full rank, Lemma 10 implies that such a Hermitian ΔA exists if

$$(V_{k+1}\underline{G}_k)^H(V_k - (A - \sigma I)V_{k+1}\underline{G}_k) = (V_{k+1}\underline{G}_k)^H V_k - (V_{k+1}\underline{G}_k)^H(A - \sigma I)V_{k+1}\underline{G}_k$$

is Hermitian. Since the first term on the right hand side is Hermitian by assumption, this is easily seen to be the case. Also by Lemma 10, ΔA is bounded by

$$\begin{aligned}\|\Delta A\| &\leq 2\|(A - \sigma I)V_{k+1}(\underline{H}_k - \underline{G}_k) - F_k\|/\sigma_{\min}(V_{k+1}\underline{G}_k)\\ &\leq 2(\|(A - \sigma I)\|_2\|V_{k+1}\|\|\underline{H}_k - \underline{G}_k\| + \|F_k\|)/\sigma_{\min}(V_{k+1}\underline{G}_k),\end{aligned}$$

and is of rank at most $2k$. □

Remark 6 If $A + \Delta A - \sigma I$ is singular, then we can use the second part of Lemma 7 to find a *Hermitian* backward error $\Delta\widetilde{A}$ arbitrarily close to ΔA such that $A + \Delta\widetilde{A} - \sigma I$ is invertible.

In order to obtain a small Hermitian backward error, we need to find a matrix $\underline{G_k}$ close to $\underline{H_k}$ such that $V_k^H V_{k+1}\underline{G_k}$ is Hermitian. One possibility is

$$\underline{G_k} := R_{k+1}^{-1}\begin{bmatrix} T_k \\ h_{k+1,k}e_k^T \end{bmatrix} R_k, \tag{22}$$

where R_k, R_{k+1} are upper triangular QR factors of V_k, V_{k+1}, respectively, and T_k is the real symmetric tridiagonal matrix with $t_{j+1,j} = t_{j,j+1} = h_{j+1,j}$ and $t_{j,j} = \Re(h_{jj})$. Then $\underline{G_k}$ is Hessenberg and computing Ritz pairs is particularly easy: we need to find vectors z and scalars μ such that

$$V_k^H(A + \Delta A - \sigma I)^{-1}V_k z = \mu V_k^H V_k z.$$

Here we have used Remark 6 in order to ensure that $A + \Delta A - \sigma I$ is invertible. By using the Krylov relation $(A + \Delta A - \sigma I)^{-1}V_k = V_{k+1}\underline{G_k}$ we obtain

$$V_k^H V_{k+1}\underline{G_k} z = \mu V_k^H V_k z.$$

Inserting the QR factorizations $V_j = Q_j R_j$, $j = k, k+1$ and the formula for $\underline{G_k}$ shown in (22) yields

$$R_k^H [I\ 0] R_{k+1} R_{k+1}^{-1}\begin{bmatrix} T_k \\ h_{k+1,k}e_k^T \end{bmatrix} R_k z = \mu R_k^H R_k z,$$

which simplifies to $T_k\tilde{z} = \mu\tilde{z}$ where $\tilde{z} = R_k z$. So, the Ritz values are just the eigenvalues of T_k (which are real, since T_k is Hermitian). To obtain the Ritz vectors, we would have to multiply $\tilde{z}$ with R_k^{-1}. However, since R_k is close to the identity matrix if the orthogonalization has been done properly (for instance, by using MGS with reorthogonalization) we can approximate $\tilde{z}$ by z. Thus, (approximations of) Ritz pairs for the choice (22) of $\underline{G_k}$ can be obtained without computing R_k, R_{k+1}. We also note that choosing the eigenpairs of T_k to construct Ritz pairs is what is done in practice.

5.3 Conditions for breakdown

We now discuss how to derive a sensible breakdown criterion based on our error analysis. We saw in Sect. 1.1 that the computed quantities V_{j+1} and $\underline{H_j}$ satisfy

$$(A - \sigma I)^{-1}(V_j + F_j) = V_{j+1}\underline{H_j}.$$

This recurrence can be rewritten as

$$(A - \sigma I)^{-1}(V_j + \widetilde{F}_j) = V_j H_j,$$

where $\widetilde{F}_j = F_j - (A - \sigma I)h_{j+1,j}v_{j+1}e_j^T$. Note that the first $j-1$ columns of $\widetilde{F}_j$ and F_j are identical. For the last column, we have

$$\widetilde{f}_j = r_j - (A - \sigma I)(g_j + h_{j+1,j}v_{j+1}),$$

where r_j is the residual from the linear system and g_j is associated column error from the orthonormalization. It is natural to declare breakdown when the error introduced by neglecting $h_{j+1,j}$ is of the same order as the errors that are present in the computation. This leads us to the following breakdown condition:

$$h_{j+1,j} < \|g_j\| + \|r_j\|/\|(A - \sigma I)v_{j+1}\|.$$

We can simplify this condition by replacing $\|g_j\|$ with its bound in (10). This yields

$$h_{j+1,j} < \eta(n, j)\|w_j\|u + \|r_j\|/\|(A - \sigma I)v_{j+1}\|. \tag{23}$$

We now discuss how to evaluate (23) in practice. If $h_{j+1,j} < \eta(n, j)\|w_j\|u$, then we can declare breakdown without further work. Otherwise we have to take the second term in (23) into consideration. If an iterative linear system solver that guarantees a residual less than some tolerance is used, then we can substitute $\|r_j\|$ in (23) by the given tolerance. If, for example, (6) is used as a stopping condition for the linear system solver, then $\|r_j\|$ is replaced by the right hand side of (6). If the residual, or any good bound for it, is not given, then we need to compute it. This is generally the case when the linear systems are solved by a direct method. Let m be a constant such that the following forward error bound holds for an arbitrary vector x

$$\|\text{float}((A - \sigma I)x) - (A - \sigma I)x\| \leq mu\|A - \sigma I\|\|x\|.$$

If $A-\sigma I$ is given as a dense matrix, we have $m = n^{3/2}$ [12, p. 70]. For sparse matrices, m can be much smaller. The computed residual $\widehat{r}_j$ satisfies

$$\begin{aligned}\|\widehat{r}_j\| &\leq (1 + u)\|\text{float}((A - \sigma I)w_j) - v_j\| \\ &\leq (1 + u)(\|r_j\| + mu\|A - \sigma I\|\|w_j\|).\end{aligned}$$

By comparing to (3), we recognize $\|A - \sigma I\|\|w_j\|mu$ as a part of the norm of a residual associated with a computed solution with corresponding backward error mu. Thus, we can compute a satisfactory $\widehat{r}_j$ if we use an extended precision $\overline{u}$ such that $m\overline{u} < u$.

For the computation of the vector $(A - \sigma I)v_{j+1}$, we have

$$\begin{aligned}\|\text{float}((A - \sigma I)v_j) - (A - \sigma I)v_j\| &\leq mu\|A - \sigma I\|\|v_j\| \\ &\leq mu\kappa(A - \sigma I)\|(A - \sigma I)v_j\|,\end{aligned}$$

and, using the reverse triangle inequality, that

$$\|(A - \sigma I)v_j\|\big(1 - mu\kappa(A - \sigma I)\big) \leq \|\text{float}((A - \sigma I)v_j)\|.$$

Thus the norm of the computed vector is accurate enough as long as $mu\kappa(A-\sigma I) \ll 1$. If $A - \sigma I$ is so ill-conditioned that this is not satisfied, then we can use an extended precision $\overline{u}$ such that $m\overline{u}\kappa(A - \sigma I) \ll 1$.

If (6) and (23) hold, then

$$\|\widetilde{f}_j\| \leq 2\big(\|v_j\|\epsilon_1 + \|A - \sigma I\|\|w_j\|(\epsilon_2 + \eta(n, j)u)\big).$$

By derivations similar to those leading to (13), we get

$$\|\widetilde{F}_j\| \leq 2\big(\sqrt{j}\|V_j\|\epsilon_1 + \sqrt{j}\|A - \sigma I\|\|V_{j+1}\|\|\underline{H}_j\|c_{jn}(\epsilon_2)\big). \tag{24}$$

From this we obtain the following "breakdown analogue" of Theorem 8.

Theorem 12 *Let $(A - \sigma I)^{-1}(V_j + \widetilde{F}_j) = V_j H_j$ be of full rank and assume $\widetilde{F}_j$ is bounded as in (24) and $\sqrt{j}\kappa(V_j)\epsilon_1 < 1$. Then there is a ΔA of rank at most j such that*

$$V_j = (A + \Delta A - \sigma I)V_j H_j$$

and

$$\|\Delta A\| \leq 2\sqrt{j}\|A - \sigma I\| \frac{\kappa(V_j)\epsilon_1 + \kappa(V_{j+1})\|\underline{H}_j\|/\sigma_{\min}(H_j)c_{jn}(\epsilon_2)}{1 - \sqrt{j}\kappa(V_j)\epsilon_1},$$

where $c_{jn}(\cdot)$ is given by (14).

The proof is omitted since it is essentially the same as the proof of Theorem 8. In a similar manner, we can get corresponding breakdown analogues to Corollary 9 and Theorem 11.

6 Conclusion

We have shown that a floating point implementation of the shift-and-invert Arnoldi algorithm, where errors from all steps of the computation are taken into account, yields computed quantities that satisfy an exact shift-and-invert Krylov recurrence of a perturbed matrix. Here, the word "Krylov" is used instead of "Arnoldi" since the computed basis cannot be guaranteed to be perfectly orthogonal. We saw that the condition number of the computed basis V_{k+1} plays a role in the bounds of the backward error. Further, we have seen that the norm of the backward error ΔA depends on $\kappa(\underline{H}_k)$. We have seen that large $\kappa(\underline{H}_k)$ are acceptable if the linear systems are only solved to a loose tolerance (18). Otherwise we argued that even if this condition number is large, the restriction to the most important part of the recurrence (that is, what is left after purging the small eigenvalues of H_k) can have a small backward error.

For Hermitian matrices A, we have shown that there is a Hermitian backward error ΔA such that the computed basis, that is, the columns of V_{k+1}, spans a Krylov subspace associated with $A + \Delta A$. However, as in the case of standard Arnoldi [15], the small

$(k+1) \times k$ matrix associated this subspace is generally not the computed Hessenberg matrix.

Finally, we noted that our error analysis yields a sensible condition for when to declare breakdown. If this condition is met, we could derive a new set of backward error bounds, which show that an invariant subspace of a perturbed matrix has been found.

References

1. Abdelmalek, N.N.: Round off error analysis for Gram–Schmidt method and solution of linear least squares problems. BIT **11**(4), 345–368 (1971)
2. Arioli, M., Duff, I., Ruiz, D.: Stopping criteria for iterative solvers. SIAM J. Matrix Anal. Appl. **13**(1), 138–144 (1992)
3. Arioli, M., Fassino, C.: Roundoff error analysis of algorithms based on Krylov subspace methods. BIT **36**(2), 189–206 (1996)
4. Arnoldi, W.E.: The principle of minimized iterations in the solution of the matrix eigenvalue problem. Q. Appl. Math. **9**, 17–29 (1951)
5. Björck, Å.: Solving linear least squares problems by Gram–Schmidt orthogonalization. BIT **7**(1), 1–21 (1967)
6. Braconnier, T., Langlois, P., Rioual, J.: The influence of orthogonality on the Arnoldi method. Linear Algebra Appl. **309**(1–3), 307–323 (2000)
7. Brent, R., Percival, C., Zimmermann, P.: Error bounds on complex floating-point multiplication. Math. Comp. **76**(259), 1469–1481 (2007)
8. Drkošová, J., Greenbaum, A., Rozložník, M., Strakoš, Z.: Numerical stability of GMRES. BIT **35**(3), 309–330 (1995)
9. Duff, I.S., Erisman, A.M., Reid, J.K.: Direct Methods for Sparse Matrices. Oxford University Press, New York (1986)
10. Giraud, L., Gratton, S., Langou, J.: Convergence in backward error of relaxed GMRES. SIAM J. Sci. Comput. **29**(2), 710–728 (2007)
11. Grcar, J.F.: Operator coefficient methods for linear equations. Technical report SAND89-8691, Sandia National Labratories (1989)
12. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2002)
13. Hochstenbach, M.E.: Probabilistic upper bounds for the matrix two-norm. J. Sci. Comput. **57**(3), 464–476 (2013)
14. Jeannerod, C.P., Rump, S.M.: Improved error bounds for inner products in floating-point arithmetic. SIAM J. Matrix Anal. Appl. **34**(2), 338–344 (2013)
15. Kandler, U., Schröder, C.: Backward error analysis of an inexact Arnoldi method using a certain Gram Schmidt variant (preprint) (2013) TU Berlin. http://www3.math.tu-berlin.de/preprints/files/Preprint-10-2013.pdf
16. Lehoucq, R.B., Meerbergen, K.: Using generalized Cayley transformations within an inexact rational Krylov sequence method. SIAM J. Matrix Anal. Appl. **20**(1), 131–148 (1998)
17. Lehoucq, R.B., Sorensen, D.C., Yang, C.: ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods. Society for Industrial and Applied Mathematics, Philadelphia (1998)
18. Mackey, D.S., Mackey, N., Tisseur, F.: Structured mapping problems for matrices associated with scalar products part I: Lie and Jordan algebras. SIAM J. Matrix Anal. Appl. **29**(4), 1389–1410 (2007)

19. Meerbergen, K., Morgan, R.: Inexact methods (section 11.2). In: Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.) Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. Society for Industrial and Applied Mathematics, Philadelphia (2000)
20. Rigal, J.L., Gaches, J.: On the compatibility of a given solution with the data of a linear system. J. Assoc. Comput. Mach. **14**(3), 543–548 (1967)
21. Simoncini, V., Szyld, D.B.: Theory of inexact Krylov subspace methods and applications to scientific computing. SIAM J. Sci. Comput. **25**(2), 454–477 (2003)
22. van der Sluis, A.: Condition numbers and equilibrium of matrices. Numer. Math. **14**(1), 14–23 (1969)
23. Smoktunowicz, A., Barlow, J.L., Langou, J.: A note on the error analysis of classical Gram–Schmidt. Numer. Math. **105**(2), 299–313 (2006)
24. Stewart, G.: A Krylov–Schur algorithm for large eigenproblems. SIAM J. Matrix Anal. Appl. **23**(3), 601–614 (2002)
25. Trefethen, L.N., Bau, D.: Numerical Linear Algebra. Society for Industrial and Applied Mathematics, Philadelphia (1997)
26. Watkins, D.: The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods. Society for Industrial and Applied Mathematics, Philadelphia (2007)